Ecological Indicators 117 (2020) 106609

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Automated detection and classification of birdsong: An ensemble approach

Stuart A. Brooker[a,*], Philip A. Stephens[a], Mark J. Whittingham[b], Stephen G. Willis[a,*]

[a] Department of Biosciences, Durham University, South Road, Durham DH1 3LE, United Kingdom
[b] School of Natural and Environmental Sciences (SNES), Agriculture Building, Newcastle University, Newcastle-upon-Tyne NE1 7RU, United Kingdom

ARTICLE INFO

Keywords:
Automated detection
Bioacoustics
Birdsong
Dawn chorus
Ensemble forecasting
Survey method

ABSTRACT

The avian dawn chorus presents a challenging opportunity to test autonomous recording units (ARUs) and associated recogniser software in the types of complex acoustic environments frequently encountered in the natural world. To date, extracting information from acoustic surveys using readily-available signal recognition tools ('recognisers') for use in biodiversity surveys has met with limited success. Combining signal detection methods used by different recognisers could improve performance, but this approach remains untested. Here, we evaluate the ability of four commonly used and commercially- or freely-available individual recognisers to detect species, focusing on five woodland birds with widely-differing song-types. We combined the likelihood scores (of a vocalisation originating from a target species) assigned to detections made by the four recognisers to devise an ensemble approach to detecting and classifying birdsong. We then assessed the relative performance of individual recognisers and that of the ensemble models. The ensemble models out-performed the individual recognisers across all five song-types, whilst also minimising false positive error rates for all species tested. Moreover, during acoustically complex dawn choruses, with many species singing in parallel, our ensemble approach resulted in detection of 74% of singing events, on average, across the five song-types, compared to 59% when averaged across the recognisers in isolation; a marked improvement. We suggest that this ensemble approach, used with suitably trained individual recognisers, has the potential to finally open up the use of ARUs as a means of automatically detecting the occurrence of target species and identifying patterns in singing activity over time in challenging acoustic environments.

1. Introduction

Autonomous recording units (ARUs) are increasingly used to gather ecological data for a diverse array of sound-producing animal taxa, including insects, anurans, cetaceans, bats, primates and birds (Sugai et al., 2019). Used appropriately, ARUs provide an efficient, standardised and unbiased data-collection procedure at lower cost than traditional site visits by skilled observers (e.g. Zwart et al, 2014). They can be deployed in situ for extended periods, recording multiple species at multiple sites simultaneously, accumulating data on spatio-temporal scales of ecological consequence, whilst limiting disturbance and reducing potentially costly visits to distant and hard-to-access locations (Blumstein et al., 2011). However, in common with other automated data collection methods in ecology (e.g. camera-traps; Norouzzadeh et al., 2018), the rate-limiting step in biodiversity studies using such data, is that of extracting information from the considerable datasets amassed. This can involve manually browsing many hours of sound recordings on spectrograms, which is a laborious task (Sebastián-González et al., 2015), potentially requiring costly teams of sound analysts (e.g. Furnas and Callas, 2015; Sanders and Mennill, 2014). Automated computer-aided signal recognition systems provide a potential solution to the problem, and reliable systems will be crucial to the viability of long-term, large-scale ecological studies using ARUs (Blumstein et al., 2011). However, despite progress in recent years, the performance of signal recognition systems has failed to keep pace with advances in acoustic data collection and storage (Wimmer et al., 2013).

The process of automatically detecting and classifying birds from sound recordings potentially presents a greater challenge than for other taxa (Brandes, 2008; Briggs et al., 2012), as bird vocalisations are typically produced within a busy sonic environment (as opposed to the ultra- or infra-sonic environments utilised by e.g. bats and cetaceans), and are prone to masking from biophony (e.g. other birds and insects), geophony (e.g. wind, rain and running water) and anthrophony (e.g. road-traffic noise and engines). Furthermore, their songs are extremely varied and complex, and when multiple species and individuals sing simultaneously, such as during the dawn chorus, elements of song overlap in time, frequency and amplitude (Luther and Wiley, 2009; Priyadarshani et al., 2018). Consequently, sound recordings made

* Corresponding authors.
E-mail addresses: s.a.brooker@durham.ac.uk (S.A. Brooker), s.g.willis@durham.ac.uk (S.G. Willis).

https://doi.org/10.1016/j.ecolind.2020.106609
Received 20 February 2020; Received in revised form 29 May 2020; Accepted 3 June 2020

during the dawn chorus often prove overwhelming for signal recognition systems, which then fail to interpret the species-specific vocalisations accurately. Species recognition during the dawn chorus has previously been identified as a particularly challenging research problem for automated detection systems (e.g. Duan et al., 2013). Yet, best practice for traditional bird surveys in many parts of the world is to survey at or around dawn during the main breeding period, to maximise the number of species detected per unit time (Bibby et al., 2000). Whether a researcher wishes to detect the presence of a species of interest, or to build a comprehensive list of species for biodiversity assessment, for most species, the dawn chorus presents the optimal opportunity to achieve this efficiently. This also applies to surveys using ARUs, especially given that battery life and data storage are often a limitation (Burivalova et al., 2019). Furthermore, singing behaviour at dawn often provides insight into breeding stage (Zhang et al., 2015), fitness of individuals (Poesel and Kempenaers, 2000) and time and energy budgets in birds (McNamara et al., 1987).

Despite the difficulties, numerous methods have been developed for automated detection and classification of birdsong (see Blumstein et al., 2011; Priyadarshani et al., 2018; Stowell and Plumbley, 2011 for reviews), drawing upon research expertise in the fields of mathematics, computer engineering, bioinformatics, acoustics and audio signal processing. As a result, several sound analysis software packages have been developed that include general-use automated signal recognition tools (hereafter 'recognisers') aimed at facilitating the use of ARUs by ecologists with only a limited understanding of the complexities of analysing acoustic data (e.g. Charif et al., 2010; Katz et al., 2016). To date, recognisers appear under-utilised in the ecological literature, and studies that have used them effectively typically document habitat occupancy and rudimentary activity levels at limited spatio-temporal scales. Many of these studies also focus on detection of distinctive, diagnostic or uncomplicated vocalisations at times of day when masking from background noise is low (e.g. Abrahams and Denny, 2018; Knight et al., 2017; Swiston and Mennill, 2009; Zwart et al., 2014). Researchers attempting more ambitious usage, such as detecting and recognising passerine songs at numerous and varied locations, have been unable to create recognisers that are fit for purpose (e.g. Sidie-Slettedahl et al., 2015). Manual scanning of spectrograms remains the best option if an accurate account of singing activity, or detection of multiple species, is required (Joshi et al., 2017; Knight et al., 2017; Sanders and Mennill, 2014; Shonfield and Bayne, 2017; Swiston and Mennill, 2009). Although recognisers are designed to facilitate signal recognition by reducing the time required to analyse large datasets, they do not fully automate the process (Charif, et al., 2010; Shonfield and Bayne, 2017). The procedure therefore, invariably involves manual verification of the detections returned, which, in itself, can be a prohibitive task.

For researchers to have confidence in the output returned, recognisers must maximise the ratio of true-positive (TP) detections over false-positive (FP) errors; to assume (likely) absence, they must eliminate FP errors entirely. To assist in this, many recognisers assign a score value to each detection, which can be taken as a confidence measure of how well the detection matches the target signal (Knight et al., 2017). In theory, higher scoring detections are more likely to originate from the target species. Many recogniser interfaces allow the user to set a score threshold, and if signals are assigned a score below the threshold, they will be omitted from the list of detections returned. Setting the threshold to a high value has the desirable effect of reducing the number of FP errors and increasing the number of true-negatives (TN), but the trade-off is fewer TP detections and more false-negative (FN) errors. Setting the threshold low will have opposite effects. Given such trade-offs, and the imperfection of recogniser performance, the score thresholds set are often based upon trial-and-error and will ultimately depend upon the question being addressed or the priorities of the research (Katz et al., 2016; Knight et al., 2017).

Despite their limitations, the various methods of signal detection amongst different recognisers may each have particular strengths when applied to certain situations and song-types, such that a combination of methods could produce a more robust and universal recogniser tool. Indeed, in both ecological studies and more widely, it is acknowledged that if individual predictive techniques provide some independent information, a combination of techniques will yield lower mean error than any one in isolation (Araújo and New, 2007). Here, we combine the performance of recognisers from four sound analysis software packages by using the scores assigned to detections to construct an ensemble model. The performance of our ensemble is compared to that of each of the recognisers in isolation in its ability to detect and classify birdsong correctly within noisy recordings made during the dawn chorus. We repeat this for five common British woodland bird species, which, together, exhibit a wide variation in song structure. Our goal is to evaluate which approaches to automated identification perform best and to test whether combining different recognisers can enhance performance across multiple species. We evaluate methods in terms of increasing TPs and decreasing, or negating, FPs, with the aim of producing a generic approach that could be more widely applied.

2. Materials and methods

2.1. Study species

Stowell and Plumbley (2011) recognise five broad song-types amongst British birds. To ensure that we tested recognisers over a varied range of songs, we used an example species with song comparable to each of these five song-type groups as follows: 1) chiffchaff *Phylloscopus collybita* (bi-syllabic), 2) wren *Troglodytes troglodytes* (few syllables, with a strong bigram structure), 3) robin *Erithacus rubecula* (large vocabulary), 4) carrion crow *Corvus corone* (less-tonal), and 5) woodpigeon *Columba palumbus* (low-pitched non-passerine).

2.2. Data collection

We collected acoustic data using ARUs (Song Meter 2+; Wildlife Acoustics Inc, Maynard, USA) mounted on tree trunks c.4 m from ground-level at each of 20 semi-natural mixed deciduous woodland study sites throughout Great Britain (Fig. 1). Each ARU was fitted with two omni-directional all-weather microphones (SMX-II; Wildlife Acoustics Inc, Maynard, USA) with a typical sensitivity of −35 to −43 dBV/pa and a frequency response of 20 Hz–20,000 Hz (Sebastián-González et al., 2015; Turgeon, et al., 2017). Recordings were made in stereo, with a sample rate of 16000 Hz and 16-bit encoding. No high-pass or bandwidth filters were applied. ARUs were configured with the respective site co-ordinates and programmed to survey continuously for 105 min, commencing 60 min prior to local sunrise every day from March to June inclusive. These surveys were repeated for each of the three years 2014 to 2016. With the exception of chiffchaff, which was absent from five sites, the study species were ubiquitous throughout our study sites.

2.3. Test dataset

We extracted 300 samples from our full dataset of acoustic surveys of the dawn chorus using stratified random sampling, ensuring that samples were evenly distributed amongst all 20 study sites (15 per site) and across all three years. Samples including persistent heavy rain and strong winds were excluded from the test dataset and substituted with a new, randomly-generated sample. A randomly selected 300 s block of time was then extracted from each of the 300 samples. The final test dataset comprised 1500 min (300 × 300 s) of acoustic survey.

2.4. Manual song detection

The test dataset was manually analysed by a single experienced ornithologist (SB), who listened to each 300 s sample whilst

Fig. 1. Locations of study sites (filled black circles). One autonomous recording unit (ARU) was installed at each site. All sites consisted of semi-natural mixed deciduous woodland habitat.

simultaneously viewing its spectrogram, and recorded all the singing events by each study species in turn (see Appendix A for definitions of singing events). Behavioural Observation Research Interactive Software (BORIS; Friard and Gamba, 2016) was used to record the timing of singing events. We used the 'live' setting on this program whilst simultaneously viewing spectrograms on Raven Pro v1.4 sound analysis software (Cornell lab of Ornithology, Ithaca, USA). If a song could not be reliably assigned to a study species (too faint/quiet, too blurred, masked by other calls, or otherwise undecipherable) it was excluded from the analysis, as were vocalisations other than song (e.g. contact and flight calls).

2.5. Automated song detection

2.5.1. Training dataset

We created individual recognisers for each study species from each of four sound analysis software packages, using singing events taken from a standardised training dataset. The full training dataset consisted of one manually-selected 105 min acoustic survey of the dawn chorus from each of our 20 study sites, or, in the case of chiffchaff, from each study site that the species was present. This ensured that the song of each study species was represented with examples of varying structure and quality; thus, creating recognisers designed for general use across multiple study sites. Acoustic surveys included within the test dataset (see section 2.3) were exempt from selection for the training dataset. Here, we provide a brief description of each of the four sound analysis software packages and the recogniser tools. A detailed methodology for recogniser construction is provided in Appendix B.

2.5.2. monitoR

MonitoR (Hafner and Katz, 2018b) is an R package offering two template-matching systems for automated detection of acoustic signals: cross-correlation and binary-point matching. We used the former for our analyses, as this method performed best in preliminary tests with our dataset. Template matching is a process in which a template of a target species' song is repeatedly scored for similarity against a moving window of an acoustic survey. MonitoR provides a score, based upon Pearson's correlation coefficient (Pearson's r), representing a detected signal's similarity to the template. We built recognisers following instructions in the demonstration vignette (Hafner and Katz, 2018a) and the suggested workflow in Katz et al. (2016).

2.5.3. Raven pro

Raven Pro (v1.4) (hereafter 'Raven') sound analysis software offers two methods for automated signal detection: a band limited energy detector, and an amplitude detector. We used the former for our analyses, as preliminary tests showed that this performed better with our dataset. The band limited energy detector operates by estimating the background noise of a signal, and uses this to find sections of the signal that exceed a user-specified signal-to-noise ratio (SNR) threshold within a specified frequency band, and during a specified time. Raven assigns an 'Occupancy' measurement to detections, which represents the percentage of samples within a selection that must exceed the background noise SNR threshold in order for the signal to be considered a positive detection. We used this measurement as a score. Raven offers a large repertoire of additional measurements applicable to detections; we selected the 'Average Power (dB)' measurement, as we surmised that this value could also predict the probability that detections are correct. We built recognisers following instructions available within the Raven v1.4 User's Manual (Charif, et al., 2010).

2.5.4. Song Scope

Song Scope (v4.1.3A; Wildlife Acoustics Inc, Maynard, USA) uses complex signal processing algorithms based upon Hidden Markov Models (HMMs) to construct recognisers from a training dataset of target vocalisations. The algorithms examine the spectral and temporal features of individual syllables and how they are organised into song. Song Scope assigns both a 'Score' and 'Quality' value to detections. Score represents the statistical fit of the detection to the recogniser's model, and Quality indicates a signal quality confidence. Detections must reach both a user-defined minimum Score and minimum Quality to count. Song Scope also returns a 'Level (dB)' value, which is the peak signal level of the vocalisation in detections. We built recognisers following instructions available within the Song Scope v4.0 documentation (Wildlife Acoustics Inc, 2011), whilst also consulting Agranat (2009) for additional advice on settings.

2.5.5. Kaleidoscope pro

Like Song Scope, Kaleidoscope Pro (v5.1.2; Wildlife Acoustics Inc, Maynard, USA) (hereafter 'Kaleidoscope') uses HMMs to detect and classify groups of syllables based upon their spatio-temporal properties, and how they combine to form phrases or song. It then groups detections into clusters based upon their similarity. Kaleidoscope assigns a score to detections, based upon their distance from the centre of the cluster. In this case, lower scores indicate better matches to the training

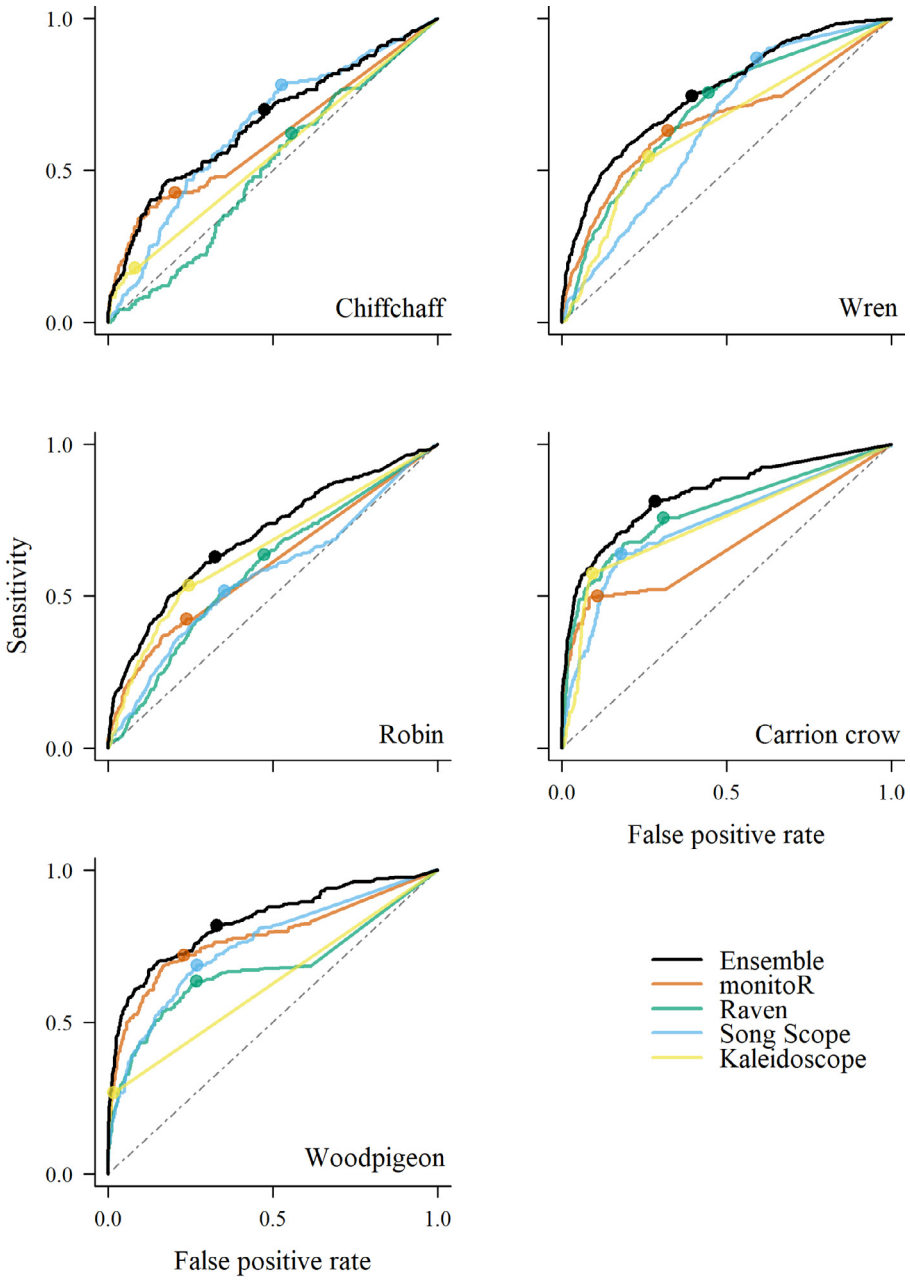

Fig. 2. The performance of an ensemble model and the four component recognisers when detecting and classifying the song of five bird species within acoustic surveys made during the dawn chorus. Filled circles show the minimum distance between the curves and the point x = 0, y = 1 (roc01). Dashed lines show random performance.

data. We constructed recognisers using the cluster analysis function, following a tutorial video available from the software developer (Wildlife Acoustics Inc, 2018a) and instructions within the Kaleidoscope v5 documentation (Wildlife Acoustics Inc, 2018b).

2.6. Ensemble model

We used the scores assigned to the detections made by the four recognisers, plus additional measurements provided by Raven (Power) and Song Scope (Quality and Level), to construct an ensemble model for each study species. We used generalised linear models (GLMs) with binomial errors, implemented using the glm function in the *stats* package in R (v3.5.2; R Core Team, 2018), to predict whether the study species was singing or not within each acoustic survey segment of 1 s duration (hereafter 'segment'), with the recogniser scores and additional measurements, and their interaction terms, as explanatory variables. We used the R package *MuMIn* (Barton, 2018) to rank candidate models by Akaike information criterion (AIC), and selected the highest ranking model. This process was repeated using cloglog, logit and probit links; the link that produced the highest ranking model with the lowest AIC was retained (Burnham and Anderson, 2002). To assess the performance of individual recognisers in isolation, binomial GLMs were used to model the probability of obtaining positive detections but only including the recogniser scores from an individual recogniser in three cases (monitoR, Raven and Kaleidoscope), or, for Song Scope, with both Score and Quality as covariates. Again, all GLMs were repeated using cloglog, logit and probit links, and the links that produced the models with the lowest AIC were chosen. See Appendix C for further details on the modelling process, and Table C.1 for model specifications.

2.7. Recogniser performance analysis

To assess the respective performances of each recogniser and the ensemble, we used area under the receiver operating characteristic curve (AUC-ROC). AUC-ROC was calculated for each species using the R package *ROCR* (Sing et al., 2005), and curves were drawn using the *PRROC* package (Grau et al., 2015). We then calculated i) the minimum distance between the ROC curves and x = 0, y = 1 (roc01), and ii) the minimum modelled probability of obtaining a positive detection at which the false positive rate (FPR) remained at zero (i.e. the probability threshold that negated FP errors but which returned TP detections), for each recogniser and the ensemble, using the R package *cutpointr* (Thiele, 2018). See Appendix D for further detail on this process.

We tested for statistical difference amongst the recognisers and the ensemble in i) AUC-ROC, and ii) roc01 using linear mixed-effects models (LMMs) implemented in the R package *lme4* (Bates et al., 2015) with model fit by maximum likelihood. AUC-ROC and roc01 performance varied amongst the study species; hence, species was included as a random intercept term in both models. We performed Tukey post-hoc pairwise tests of recognisers using the R package *emmeans* (Length et al., 2019). We confirmed that normality and homoscedasticity assumptions were met by plotting the model residuals as Q-Q plots and against fitted values respectively.

To test the ensemble's ability to recognise broad-scale patterns in singing activity over time, we applied the ensemble models to the 60 × 300 s acoustic surveys selected to be the model test data (see Appendix C), omitting samples from which the study species was absent (chiffchaff was excluded from this analysis due to the low number of datapoints following these omissions), and including all 300 segments of those that remained. We used the roc01 probabilities (as defined above) as cutpoints, and correlated the number of segments within each sample survey identified as positive detections by the ensemble against the corresponding numbers identified by manual analysis. Pearson's r was calculated as a measure of similarity. To demonstrate the ensemble's potential to recognise fine-scale patterns in singing activity, we manually selected a sample survey for each species, and aggregated the segments into 30 × 10 s blocks. We then correlated the number of segments within each 10 s block identified as positive detections by the ensemble with the corresponding numbers identified by manual analysis, and calculated Pearson's r.

3. Results

The ensemble model produced higher AUC-ROC values than all four component recognisers in isolation for all study species (Fig. 2; Table 1). The ensemble also attained lower roc01 values than all component recognisers in isolation for all study species, with the exception of chiffchaff, where Song Scope attained a lower roc01 value (Fig. 2; Table 1). No one recogniser in isolation performed consistently better, in terms of AUC-ROC or roc01, than any other across all study species (Fig. 2; Table 1). The sensitivity (i.e. the proportion of study species' 1 s singing events correctly identified as such) of the ensemble model at the optimal (lowest) roc01 cutpoint value for each study species averaged 74% amongst the species (chiffchaff = 70%, wren = 74%, robin = 63%, carrion crow = 81% and woodpigeon = 82%; Fig. 2), whilst sensitivity averaged across all component recognisers and study species at their respective optimal roc01 cutpoint values was 59% (chiffchaff = 50%, wren = 70%, robin = 53%, carrion crow = 62% and woodpigeon = 58%). These sensitivity values, however, were achieved at a cost of varying FPRs (Table D.1). The ensemble returned a minimum probability of obtaining a positive detection, whilst suppressing FPR to zero, for all study species. At this probability, FP errors were negated whilst TP detections remained. No recogniser in isolation achieved this for all study species, and Raven did not achieve this for any (Table D.2).

AUC-ROC was significantly different amongst the recognisers and the ensemble ($\chi^2(4) = 57.63$, $p < 0.001$). Tukey post-hoc tests showed that the ensemble attained significantly higher AUC-ROC than did all recognisers in isolation (monitoR, $p = 0.020$; Raven, $p < 0.001$; Song Scope, $p < 0.001$; Kaleidoscope, $p < 0.001$). Additionally, both monitoR ($p < 0.001$) and Raven ($p = 0.030$) attained significantly greater AUC-ROC than Kaleidoscope (Fig. 3a). Likewise, the roc01 statistic was significantly different amongst the recognisers and the ensemble ($\chi^2(4) = 112.63$, $p < 0.001$). Tukey post-hoc tests again showed that the performance of the ensemble was significantly better, with roc01 less than that of all other recognisers in isolation (monitoR, $p < 0.001$; Raven, $p = 0.036$; Song Scope, $p < 0.001$; Kaleidoscope, $p < 0.001$). Additionally, the roc01 of monitoR ($p < 0.001$), Raven ($p < 0.001$) and Song Scope ($p < 0.001$) were all significantly lower than that of Kaleidoscope (Fig. 3b).

The number of segments within sample surveys identified by the ensemble as positive singing events correlated positively with the numbers identified by manual analysis for all species tested (Fig. 4). Pearson's r was moderate for three study species, and strong for carrion crow (Fig. 4). Likewise, the number of segments within 10 s blocks of chosen sample surveys identified by the ensemble as positive singing events, correlated positively with the numbers identified by manual analysis for all study species. Pearson's r ranged from weak (chiffchaff) to very strong (robin; Fig. 5).

4. Discussion

If automated acoustic recognisers are to be more widely adopted in ecological studies, there is a need for improved recogniser performance in detecting and classifying vocalisations within noisy acoustic surveys. We assessed the individual performance of four readily-available recognisers and found that their ability to detect the singing events of bird species with contrasting vocalisations was highly variable. In parallel, we developed an ensemble approach, whereby scores assigned to detections made by the four recognisers were combined to model probabilities of singing events by individual species. Our ensemble model

Table 1
The area under the receiver operating characteristic curve (AUC-ROC) and the minimum distance from the ROC curve and the point x = 0, y = 1 (roc01) for an ensemble model and the four component recognisers when detecting and classifying the song of five bird species within acoustic surveys made during the dawn chorus. Lower roc01 values are optimal as they represent greater sensitivity (the proportion of species' 1 s singing events correctly identified as such) relative to the corresponding false positive rate (the proportion of species' 1 s non-singing events incorrectly identified as 1 s singing events). The ensemble attained the lowest roc01 values for all species, with the exception of chiffchaff, where Song Scope roc01 was lowest.

Species	Ensemble		monitoR		Raven		Song Scope		Kaleidoscope	
	AUC-ROC	roc01	AUC-ROC	roc01	AUC-ROC	roc01	AUC-ROC	roc01	AUC-ROC	roc01
Chiffchaff	0.658	0.528	0.606	0.616	0.502	0.689	0.640	0.496	0.548	0.829
Wren	0.756	0.412	0.661	0.472	0.696	0.442	0.647	0.479	0.641	0.523
Robin	0.699	0.476	0.604	0.633	0.591	0.588	0.570	0.608	0.653	0.525
Carrion crow	0.836	0.268	0.669	0.512	0.782	0.337	0.738	0.393	0.739	0.433
Woodpigeon	0.832	0.291	0.779	0.334	0.676	0.437	0.753	0.385	0.626	0.733

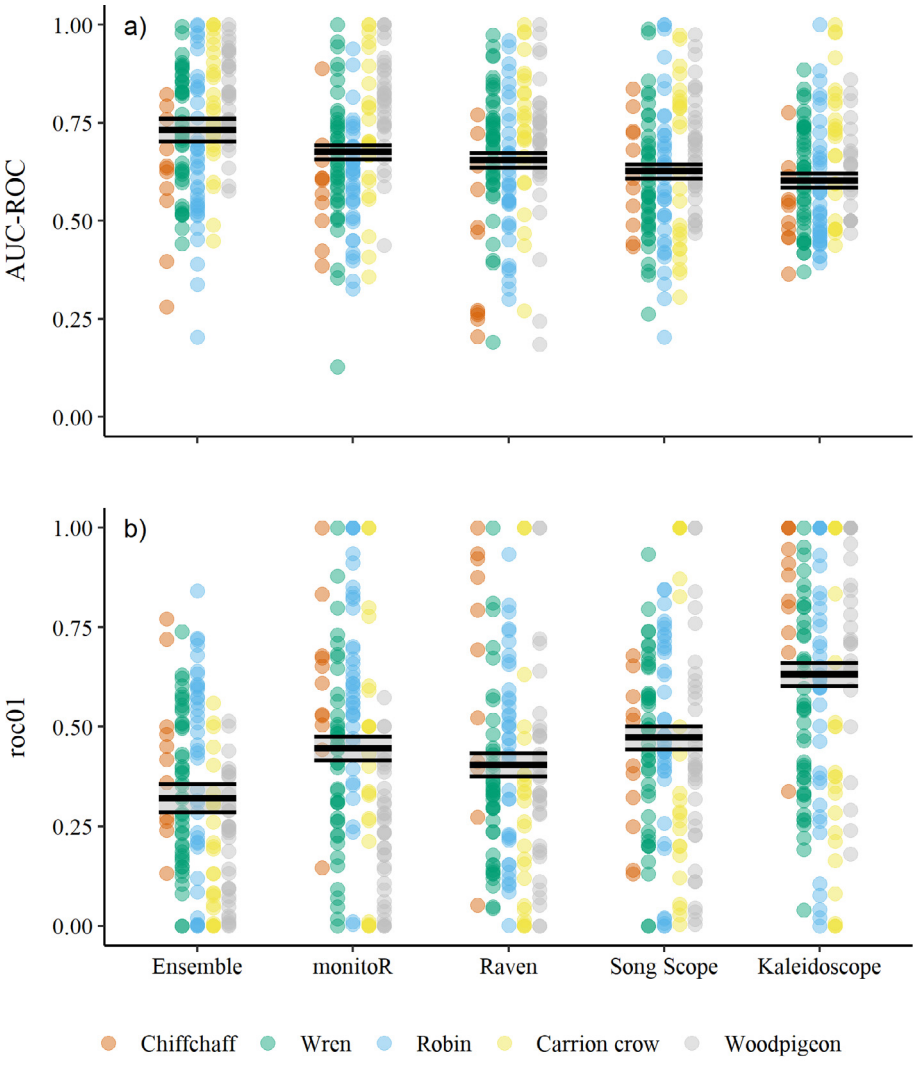

Fig. 3. Results of linear mixed-effects models (LMMs) testing for difference in a) the area under the receiver operating characteristic curve (AUC-ROC), and b) the minimum distance between the ROC curve and x = 0, y = 1 (roc01) of an ensemble model and the four component recognisers when detecting and classifying the song of five bird species within acoustic surveys made during the dawn chorus. Lower roc01 values are optimal as they represent greater sensitivity relative to the corresponding false positive rate. Thick horizontal bars represent mean values, and thin bars represent ± 1SE, having accounted for the random intercept effect of species. Data points of species are plotted; n = 155 for each recogniser.

performed significantly better than all component recognisers in isolation when tested on the song of five species in acoustic surveys made during the dawn chorus at 20 woodland sites throughout Great Britain. The mean probability of the ensemble correctly identifying individual singing events across our five study species was 74%, compared to 59% probability when the respective performances of the component recognisers were averaged across the study species. The ensemble worked by 'weighting' the scores of the component recognisers, improving classification of the 'true signal', and reducing both the error and unreliability of the recognisers when operated in isolation (Araújo and New, 2007). Hence, the ensemble takes the particular strengths from each recogniser's detection method, resulting in a favourable performance across all species tested. Considering that our study species represent the five broad song-types recognised amongst British birds (cf. Stowell and Plumbley, 2011), and are likely to be representative of song-types more broadly, it is reasonable to postulate that our ensemble method would perform favourably across other bird communities and also across other taxa and regions.

The performance of the individual recognisers was inconsistent across our five study species, reflecting the suitability of their respective signal detection methods to particular song structures, frequency ranges, and background noise; no individual recogniser was comprehensive in its ability (c.f. Brandes, 2008). For example, Raven concentrates on detecting the energy within a specified frequency band, and does not consider the internal structure of a song (Duan et al., 2013). It is, therefore, prone to a high FPR. This was especially apparent with chiffchaff song, where Raven's performance was barely better than random (Fig. 2; Table 1). Despite chiffchaff song being of a relatively simple structure, Raven was unable to discriminate between the target signal and background noise in the 3.5–7 kHz frequency band in the period around sunrise, when most species participate in the chorus. Song Scope performed considerably better in this situation, despite its method of detecting song structure using HMMs also being sensitive to noise (Briggs et al., 2012; Duan et al., 2013). By contrast, Song Scope performed poorly for wren, despite wren song being delivered at the same time as chiffchaff, and in a broadly overlapping frequency band. This may be due to the high amplitude of wren song increasing the SNR, allowing Raven to detect it more easily, whereas there was sufficient variability in wren song structure across all dates and sites in the test dataset to limit discrimination by Song Scope. In deciduous woodland, wren typically sing at lower elevations (c.3 m; Holland et al., 1998) than chiffchaff (which sings high in the canopy; Rodrigues, 1996). In our study, wren song was, thus, not only closer to the ARU microphones, which were set at c.4 m, but was also on a much more similar transverse plane, which can be beneficial for sound propagation beneath the woodland canopy at dawn (Wiley and Richards, 1978). The resultant difference in amplitude might explain the difference in their detection by Raven. In a prior comparison of recogniser performance, detecting the distinctive call of common nighthawk *Chordeiles minor* in less complex acoustic conditions at twilight, Knight et al. (2017) found

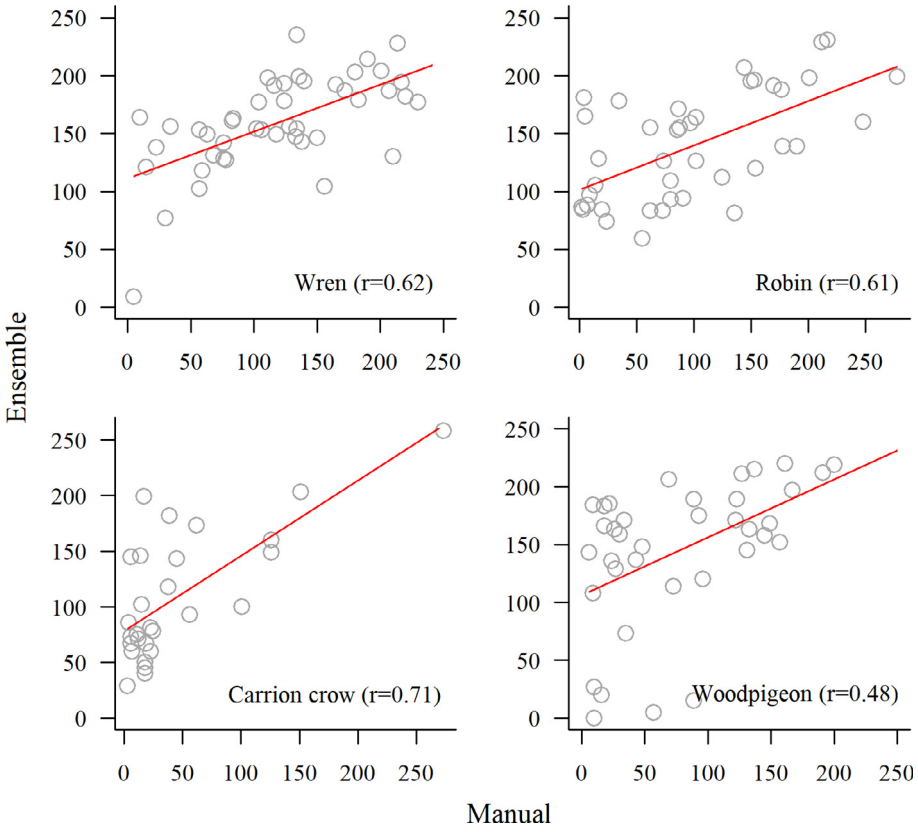

Fig. 4. The number of 1 s segments within each of n × 300 s acoustic surveys of the dawn chorus identified as positive singing events by manual analyses versus the number predicted by automated analyses using an ensemble recogniser model, and the Pearson's correlation coefficient, for four bird species with differing song-types. Wren, n = 45; robin, n = 40; carrion crow, n = 28; woodpigeon, n = 37.

that Song Scope performed best (as measured using AUC-ROC), followed by monitoR, then Raven, then Kaleidoscope. From our study species, carrion crow most closely resembles common nighthawk in terms of the structure of its vocalisation, but we obtained a very different order in AUC-ROC performance, i.e. Raven then Kaleidoscope/Song Scope, then monitoR (Fig. 2; Table 1), highlighting the potential problems of relying on only one recogniser for detection. It is apparent, therefore, that relative performances of individual recognisers are variable, and dependent upon the species and the situation within which acoustic surveys are made, and, no doubt, in the methods applied by the user during their construction. By comparison, the relative performance of our ensemble model remained consistently high across all species tested.

In many applications of ARUs, the minimum requirement is detecting the occurrence, or occupancy probability, of a species of interest at a given location (e.g. Furnas and Callas, 2015). Unfortunately, recognisers invariably return FP errors from acoustic surveys, which are particularly problematic when the species of interest is absent from the location, and which contravene a major assumption of many occupancy models (MacKenzie et al., 2006). This error can be reduced, or resolved, if there is a minimum probability of obtaining a positive detection at which the FPR remains at zero. We showed that none of the recognisers tested in isolation could achieve this probability cutpoint across all five of our study species (no individual recogniser enabled the detection of more than three species), but that this was achievable using our ensemble model (Table D.2). This means that to determine occurrence for each of our study species, we need only consider the detections made at or above the minimum threshold probability. Within this reduced dataset, we should be confident that the detections are of the target species only. If no detections are returned at or above the minimum probability, and the target species is otherwise a reliable contributor to the dawn chorus, we could infer that the species is absent. The more TP detections that exceed the minimum probability (Table D.2), the more confident this assumption should be. Importantly, an ensemble approach might, thus, enable the use of ARUs to determine apparent species presence-absence data for sites, if recognisers are available for all candidate species.

When accurate accounts of daily or seasonal patterns in song frequency or singing behaviour is important, a large majority of the singing events within acoustic surveys must be detected (Shonfield and Bayne, 2017), whilst FP errors remain negligible. In this regard, a good recogniser will minimise the distance from the ROC curve to the point x = 0, y = 1 (where the distance is denoted roc01). This was beyond the capabilities of all individual recognisers tested for most of our study species singing during the dawn chorus, and the ensemble also fell short for some species in its performance here (Fig. 2; Table 1). This was particularly true for chiffchaff, where sensitivity at the optimal roc01 was 70%, which was attained at the cost of a 48% FPR (Fig. 2; Table D.1). The best performing ensemble model was for carrion crow, where sensitivity at the optimal roc01 was 81% at a cost of a 28% FPR (Fig. 2; Table D.1). Nevertheless, the roc01 for the ensemble across the five study species was significantly less than for all component recognisers in isolation (Fig. 3b), and, for all study species except Song Scope's chiffchaff recogniser, the ensemble model had a lower or equal FPR for any given sensitivity value (Fig. 2).

An ensemble approach based on the best available current recognisers is still only partly capable of correctly detecting and classifying all individual singing events of species. In particular, when the singing activity of our study species in acoustic surveys was low, the ensemble had high FP rates (note high y-intercepts in Fig. 4). When the singing activity of the species was greater, the ensemble correlated well with the observed number of singing events over broad timescales (i.e. 300 s; Fig. 4). The ensemble also demonstrated potential for very high performance in recognising singing activity patterns over fine timescales (i.e. 10 s) for most species tested (Fig. 5), and although the ensemble model for chiffchaff largely failed to identify the nuances in timing of singing events, it still correctly estimated the mean number of events across the sample (Fig. 5). However, further development of

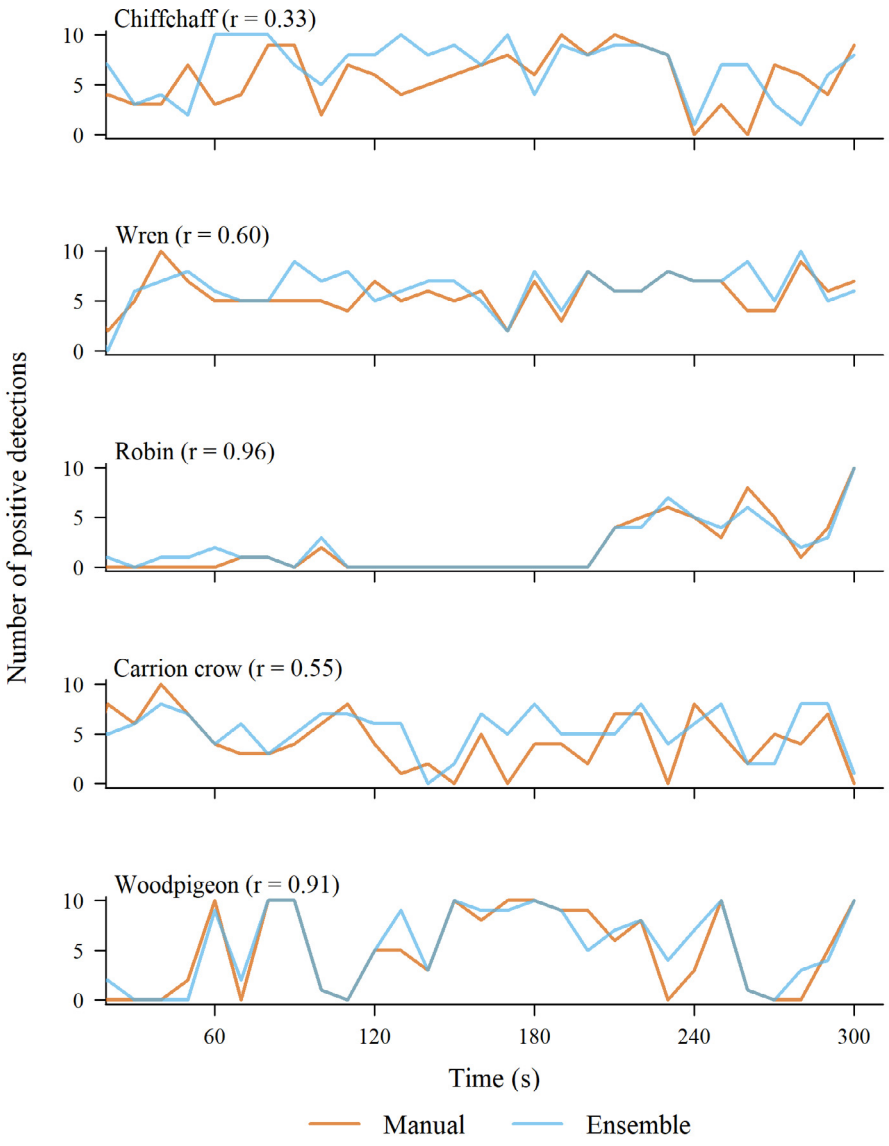

Fig. 5. The number of seconds per 10 s interval that the songs of five bird species were detected by manual analyses of sample 300 s acoustic surveys of the dawn chorus (one survey selected for each species), and the corresponding number returned by automated analyses using ensemble recogniser models. The ensemble models were capable of identifying fine-scale patterns in song output over time for the range of species. Pearson's correlation coefficients, assessing the relationship between the two methods, are shown in parentheses.

individual recognisers and the ensemble approach will be required for reliable application to studies on song output and singing behaviour.

Building ensemble recogniser models can be a lengthy process, as they require familiarity with the controls and construction of each component recogniser. Nonetheless, if they are used to examine large acoustic datasets, the enhanced performance of ensembles over the use of the component recognisers in isolation will likely out-weigh the initial time invested and, once constructed, they can be applied to a wide range of species, study sites and datasets. An alternative to investing in building an ensemble recogniser would be to allocate effort to training an individual recogniser. However, diminishing returns, together with the relatively narrow domain of performance of each individual recogniser, suggest that the outcome would be unlikely to match an ensemble approach in its breadth. Ours and previous studies suggest that major improvements can still be made to available recognisers. Future improvements to any one recogniser are also likely to improve the performance of an ensemble modelling approach, enabling a much wider utility of ARUs for ecological studies.

With both diversity and abundance of species declining at greater rates than ever before in human history (IPBES, 2019), there is a pressing need to monitor the state of our wildlife. We present a method based on five species with different vocal characteristics that improves acoustic signal recognition performance significantly. Our ensemble method offers the potential for inexpensive, robust monitoring of species. Clearly, the method needs to be tested on a wider range of species, but the potential use of ARUs for widespread use is now within our grasp. Our ensemble approach could be used for a range of purposes, including to provide evidence for: policy makers (e.g. the presence of qualifying species in protected areas), those wishing to provide evidence of the presence of species on sites notified for developments (e.g. Environmental Impact Assessments), and scientists exploring ecological and behavioural research questions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to

influence the work reported in this paper.

Acknowledgements

The authors are extremely grateful to Mrs D. Bell, Natural England, the Royal Society for the Protection of Birds (RSPB), the Wildlife Trusts, and the Woodland Trust for their kind permission to install autonomous recording units at their reserves. We thank Stuart Newson and two anonymous reviewers for their helpful comments to improve this manuscript.

Funding

This work was supported by the Natural Environment Research Council (NERC) [Grant ref: NE/L002590/1] in a CASE Partnership with the British Trust for Ornithology (BTO).

Appendix. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2020.106609.

References

Abrahams, C., Denny, M.J.H., 2018. A first test of unattended, acoustic recorders for monitoring Capercaillie Tetrao urogallus lekking activity. Bird Study 65, 197–207. https://doi.org/10.1080/00063657.2018.1446904.

Agranat, I. D., 2009. Automatic detection of cerulean warblers (pp. 1–13). Washington D.C.: USDA Forest Service. Retrieved from http://www.fs.fed.us/t-d/programs/im/acoustic_wildlife/Cerulean%20Warbler_%20Report_Final.pdf.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22, 42–47. https://doi.org/10.1016/j.tree.2006.09.010.

Barton, K., 2018. MuMIn: multi-model inference (Version 1.42.1) [R]. Retrieved from https://CRAN.R-project.org = MuMIn.

Bates, D.M., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48.

Bibby, C., J., Burgess, N., D., Hill, D., A., Mustoe, S., H., 2000. Bird census techniques, 2nd ed. Academic Press, London.

Blumstein, D.T., Mennill, D.J., Clemins, P., Girod, L., Yao, K., Patricelli, G., Deppe, J.L., Krakauer, A.H., Clark, C., Cortopassi, K.A., Hanser, S.F., McCowan, B., Ali, A., Kirschel, A.N.G., 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. J. Appl. Ecol. 48, 758–767. https://doi.org/10.1111/j.1365-2664.2011.01993.x.

Brandes, T.S., 2008. Automated sound recording and analysis techniques for bird surveys and conservation. Bird Conservation International 18, S163–S173. https://doi.org/10.1017/S0959270908000415.

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J.K., Hadley, A.S., Betts, M.G., 2012. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. The Journal of the Acoustical Society of America 131, 4640–4650. https://doi.org/10.1121/1.4707424.

Burivalova, Z., Game, E.T., Butler, R.A., 2019. The sound of a tropical forest. Science 363, 28–29. https://doi.org/10.1126/science.aav1902.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. Springer-Verlag, New York.

Charif, R. A., Waack, A. M., Strickman, L. M., 2010. Raven Pro 1.4 user's manual. Cornell Lab of Ornithology, Ithaca, New York.

Duan, S., Zhang, J., Roe, P., Wimmer, J., Dong, X., Truskinger, A., Towsey, M., 2013. Timed probabilistic automaton: a bridge between Raven and Song Scope for automatic species recognition. In Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference (pp. 1519–1524). Bellevue, Washington, USA. Palo Alto, California, USA. Retrieved from https://www.aaai.org/ocs/index.php/IAAI/IAAI13/paper/view/6092/6429.

Friard, O., Gamba, M., 2016. BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. Methods Ecol. Evol. 7, 1325–1330. https://doi.org/10.1111/2041-210X.12584.

Furnas, B.J., Callas, R.L., 2015. Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. J. Wildl. Manag. 79, 325–337. https://doi.org/10.1002/jwmg.821.

Grau, J., Grosse, I., Keilwagen, J., 2015. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics 31, 2595–2597. https://doi.org/10.1093/bioinformatics/btv153.

Hafner, S. D., Katz, J., 2018a. A short introduction to acoustic template matching with monitoR [R]. Retrieved from https://cran.r-project.org/web/packages/monitoR/vignettes/monitoR_QuickStart.pdf.

Hafner, S. D., Katz, J., 2018b. monitoR: acoustic template detection in R (Version 1.0.7) [R]. Retrieved from http://www.uvm.edu/rsenr/vtcfwru/R/?Page = monitoR/monitoR.htm.

Holland, J., Dabelsteen, T., Pedersen, S.B., Larsen, O.N., 1998. Degradation of wren Troglodytes troglodytes song: Implications for information transfer and ranging. The Journal of the Acoustical Society of America 103, 2154–2166. https://doi.org/10.1121/1.421361.

IPBES., 2019. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (Advance unedited version). IPBES. Retrieved from https://www.ipbes.net/news/ipbes-global-assessment-summary-policymakers-pdf.

Joshi, K.A., Mulder, R.A., Rowe, K.M.C., 2017. Comparing manual and automated species recognition in the detection of four common south-east Australian forest birds from digital field recordings. Emu - Austral Ornithology 117, 233–246. https://doi.org/10.1080/01584197.2017.1298970.

Katz, J., Hafner, S.D., Donovan, T., 2016. Tools for automated acoustic monitoring within the R package monitoR. Bioacoustics 25, 197–210. https://doi.org/10.1080/09524622.2016.1138415.

Knight, E., Hannah, K., Foley, G., Scott, C., Brigham, R., Bayne, E., 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology 12. https://doi.org/10.5751/ACE-01114-120214.

Length, R., Singmann, H., Love, J., Buerkner, P., Herve, M., 2019. emmeans: estimated marginal means, aka least-square means (Version 1.3.2) [R]. Retrieved from https://github.com/rvlength/emmeans.

Luther, D.A., Wiley, R.H., 2009. Production and perception of communicatory signals in a noisy environment. Biol. Lett. 5, 183–187. https://doi.org/10.1098/rsbl.2008.0733.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L., Hines, J.E., 2006. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Elsevier Academic Press, London.

McNamara, J.M., Mace, R.H., Houston, A.I., 1987. Optimal daily routines of singing and foraging in a bird singing to attract a mate. Behav. Ecol. Sociobiol. 20, 399–405. https://doi.org/10.1007/BF00302982.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J., 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. 115, E5716–E5725. https://doi.org/10.1073/pnas.1719367115.

Poesel, A., Kempenaers, B., 2000. When a bird is tired from singing: a study of drift during the dawn chorus. Etologia 8, 1–7.

Priyadarshani, N., Marsland, S., Castro, I., 2018. Automated birdsong recognition in complex acoustic environments: a review. J. Avian Biol. 49. https://doi.org/10.1111/jav.01447.

Core Team, R., 2018. R: A language and environment for statistical computing (Version v3.5.2). R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues, M., 1996. Song activity in the chiffchaff: territorial defence or mate guarding? Anim. Behav. 51, 709–716. https://doi.org/10.1006/anbe.1996.0074.

Sanders, C.E., Mennill, D.J., 2014. Acoustic monitoring of nocturnally migrating birds accurately assesses the timing and magnitude of migration through the Great Lakes. The Condor 116, 371–383. https://doi.org/10.1650/CONDOR-13-098.1.

Sebastián-González, E., Pang-Ching, J., Barbosa, J.M., Hart, P., 2015. Bioacoustics for species management: two case studies with a Hawaiian forest bird. Ecol. Evol. 5, 4696–4705. https://doi.org/10.1002/ece3.1743.

Shonfield, J., Bayne, E., 2017. Autonomous recording units in avian ecological research: current use and future applications. Avian Conservation and Ecology 12. https://doi.org/10.5751/ACE-00974-120114.

Sidie-Slettedahl, A.M., Jensen, K.C., Johnson, R.R., Arnold, T.W., Austin, J.E., Stafford, J.D., 2015. Evaluation of autonomous recording units for detecting 3 species of secretive marsh birds. Wildl. Soc. Bull. 39, 626–634. https://doi.org/10.1002/wsb.569.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940–3941. https://doi.org/10.1093/bioinformatics/bti623.

Stowell, D., Plumbley, M. D., 2011. Birdsong and C4DM: a survey of UK birdsong and machine recognition for music researchers (No. C4DM-TR-09–12, v1.2) (pp. 2–19). London: Centre for Digital Music, Queen Mary, University of London. Retrieved from http://c4dm.eecs.qmul.ac.uk/papers/2010/Stowell2010-C4DM-TR-09-12-birdsong.pdf.

Sugai, L.S.M., Silva, T.S.F., Ribeiro, J.W., Llusia, D., 2019. Terrestrial passive acoustic monitoring: review and perspectives. Bioscience 69, 15–25. https://doi.org/10.1093/biosci/biy147.

Swiston, K.A., Mennill, D.J., 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of pileated, pale-billed, and putative ivory-billed woodpeckers. J. Field Ornithol. 80, 42–50. https://doi.org/10.1111/j.1557-9263.2009.00204.x.

Thiele, C., 2018. cutpointr: Determine and evaluate optimal cutpoints in binary classification tasks (Version 0.7.4) [R]. Retrieved from https://github.com/thiele/cutpointr.

Turgeon, P., Van Wilgenburg, S., Drake, K., 2017. Microphone variability and degradation: implications for monitoring programs employing autonomous recording units. Avian Conservation and Ecology 12. https://doi.org/10.5751/ACE-00958-120109.

Wildlife Acoustics, Inc., 2011. Song Scope: bioacoustics software v4.0 documentation. Wildlife Acoustics, Inc. Maynard, USA.

Wildlife Acoustics, Inc., 2018a. Cluster analysis in-depth (Parts 1-3). Retrieved 10 August 2018, from https://www.wildlifeacoustics.com/products/kaleidoscope-pro/tutorial-videos/944-cluster-analysis.

Wildlife Acoustics, Inc., 2018b. Kaleidoscope Pro 5: User Guide (Version v5). Wildlife Acoustics, Inc. Maynard, USA.

Wiley, R.H., Richards, D.G., 1978. Physical constraints on acoustic communication in the atmosphere: implications for the evolution of animal vocalizations. Behav. Ecol. Sociobiol. 3, 69–94. https://doi.org/10.1007/BF00300047.

Wimmer, J., Towsey, M., Roe, P., Williamson, I., 2013. Sampling environmental acoustic recordings to determine bird species richness. Ecol. Appl. 23, 1419–1428. https://doi.org/10.1890/12-2088.1.

Zhang, V.Y., Celis-Murillo, A., Ward, M.P., 2015. Conveying information with one song type: changes in dawn song performance correspond to different female breeding stages. Bioacoustics 25, 19–28. https://doi.org/10.1080/09524622.2015.1076348.

Zwart, M.C., Baker, A., McGowan, P.J.K., Whittingham, M.J., 2014. The use of automated bioacoustic recorders to replace human wildlife surveys: an example using nightjars. PLoS ONE 9, e102770. https://doi.org/10.1371/journal.pone.0102770.